Dynamic filters selection for textual document images denoising

Hubert Cecotti, Abdel Belaïd

Dynamic filters selection for textual document images denoising

Hubert Cecotti
Institute of Automation (IAT), University of Bremen, Germany
cecotti@iat.uni-bremen.de

Abdel Belaïd
LORIA, Vandoeuvre-Lés-Nancy, France
abelaid@loria.fr

Abstract

For a document class, one challenge in document restoration is to automatically find a set of filters, which are adapted to the degradation level of the images. Furthermore, it is important to know what filters and where they can be applied advantageously. In this paper, we present a multi classifiers solution for the extraction of linear filters. These filters are used for binarization and image denoising. The technique starts by clustering close pixels by K-means in as many clusters as filters. Each cluster is dedicated to a filter, which corresponds to a supervised neural network. These classifiers are trained according to a binarized image that is weighted function to erosion transformation effects. The presented method is compared to classical binarization techniques in literature. Its effect on the commercial OCR performances reaches a gain from 0,16% for Finereader7 and 1,06% for Omnipage14 for the recognition rate.

1 Introduction

In image restoration, images must not be processed in an identical way. Whereas some parts are very noisy, others are clean and need no transformation. For example, a transformation on a clean part may damage characters on a textual document image. For the noisy parts, we can observe different kinds of noise and different artifacts on the characters. It can be due to a bad digitalization, a bad binarization, or simply to the age of the document. One solution consists of using a multi-classifier system (MCS) as many kinds of problems may occur. Usually, MCS are based on the outputs combination of a classifier set. Some researchers have proposed an alternative approach [4]. It is based on the concept of dynamic classifier selection (DCS). DCS is based on the definition of a function that selects, for each pattern, the most adapted and probable classifier. Anderson and Van Essen have proposed a general strategy for dynamic control of information flow between neurons clusters at different levels of the visual pathway [3]. Some means of dynamically controlling how retinal outputs map onto higher-level targets is desirable. The solution of Anderson involves "shifter circuits, which allows for dynamic shifts in the relative alignment of input and output arrays without loss of local spatial relationships. Furthermore, the dynamic shifts need a controller. In the presented system, the controller is an external classifier which assigns for each input the right shift (i.e. the right classifier). For each possible shift of the information flow, instead of considering one adaptive classifier, we consider a set of different classifiers. This classification allows determining what filters to apply and where in each image. For a very large number of pixels to process, we consider a fixed number K of deformations/noises. For all the pixels, we propose to create K solutions (i.e. K filters). This paper is organized as follows. Section 2 describes the proposed technique. Classifiers are defined in section 3. Section 4 shows how the results are evaluated. Section 5 presents the creation of the ground truth for the classifiers learning. The last section is dedicated to the experiments for document recognition with OCR.

2 The proposed approach

The proposed approach consists in defining as many filters as necessary to pre-process an image (cleaning, binarization, contrast enhancing, etc. for its different sets of pixels). The system is a trained function to an uncertain ground truth, which serves as a base for the learning. The training of the system is in two steps.

First, the training of the classifier selector. This step aims to categorize the different sets of points where filters are applied. Then, the training of the classifier function for the controller. Each classifier will behave as a filter.

The problem can be formalized as follows: let I be the set of 8 bit grey level image samples representing a document class. For each point $P(x, y)$ of I, we determine its neighborhood $V(x, y)$ composed of the points $P(x_i, y_i)$ such that $(-c \leq xi - x < c)$ and $(c \leq yi - y < c)$ where $2c + 1$ is the size of the filtering kernel. For each $V(x, y)$, a set of NF features $F(x, y)$ is extracted. If the features are just pixel values, then $V(x, y) = F(x, y)$. For our experiments, we have fixed $NF = (2c + 1)^2$. K clusters are constituted by applying the unsupervised classifier, K-means, based on the features F.

3 Classifiers training

Each classifier corresponds to the application of a linear filter, which has the same size as $V(x, y)$. The classifiers are based on a Multi-Layer Perceptron. The inputs correspond to the $V(x, y)$. They are centered between -1 and 1, without shifting or scaling. The output layer is composed of one neuron. This neuron corresponds to the desired pixel value. Let $w(i, j)_k$ be a weight linking the input pixel (i,j) to the output neuron, where $1 \leq k < K$, and $(i, j) \in \{0, , 2c\}^2$. The weights are initialized with 0 except for the center that is 1. The activation function f of each neuron is defined by $f(\sigma) = 1.7159 * tanh((2.0/3.0) * \sigma)$ in order to get an almost linear function with $f(1) = 1$ and $f(-1) = -1$, the input and the output being in the same range [6]. The state of each output neuron is $y = f(\sigma)$ where :

$$\sigma = \sum_{i=0}^{2c} \sum_{j=0}^{2c} w(i, j)_k . F(x - c + i, y - c + j)$$

For binarization purpose, the threshold is 0. At the end of the training, we have as many classifiers as filters, each one of them is representing a cluster of pixels. Usually, if the images belong to the same class, the barycenters of the clusters are similar.

4 Evaluation procedure

Several solutions are possible in order to evaluate the filtering approaches. We distinguish three methods:

- The manual evaluation. One or several experts judge the quality of the filters effects. The evaluation can be global or it can be the combination of different criteria. For example, for the binarization evaluation, Trier uses the following criteria: the broken lines structures, the broken symbols or characters, the blurring of lines symbols and text, the loss of complete objects, and the noise in homogeneous areas [11].

- The raw automatic evaluation. It can be based on just pixel level, with criteria such as the homogeneity or the contrast variation. It can also be performed by comparison with an image ground truth. This can be done by using a digital image editing software, by superposing a layer on the original image, and by rewriting the text on this layer [7, 10].

- The automatic evaluation is driven by the application. In our case, we used commercial OCR as benchmark tools. The text recognition analysis is obtained by differentiating the OCR results and a textual ground truth of the document L_{GT-TXT}. This ground truth allows differentiating the errors among: confusion, addition, deletion, fusion and cut. These errors are advantageously taken into account by the appropriate dynamic programming algorithm of Seni [9].

5 Ground truth creation

For image filtering and binarization methods, it is very awkward to define both quality measures for qualifying the performance and a ground truth. Although the performances are analyzed on the character level, it is difficult to give such results as a feedback for learning the filters. Indeed they are learnt on the pixel level. The previous section dealt with the performance evaluation relatively to L_{GT-TXT}. We propose a solution for creating an image ground truth, L_{GT-IMG}, for the classifiers learning. In our case, we consider labeled and unlabeled data where each data corresponds to a pixel on the image. The creation of the ideal document image is long and difficult: the characters must be cleaned; the noise must be removed manually. In order to avoid such task, we consider that a good binarization technique can be used as the ground truth core. We will consider some regions of the binarized image as labeled pixels whereas we fix a probability of the label for the remaining pixels.

The choice of the binarization algorithm is made with different tries on some documents. The Otsu method is presented as giving good results for historical printed documents [5, 8]. However, for our set of documents, the Sauvola method allowed achieving better results [**?**]. We do not consider the Sauvola result as the final L_{GT-IMG}. Each pixel is a weighted function to

its potential accuracy. The classifiers are not trained on a real ground truth but on partially correct information. We introduce a rule that defines the strength $\gamma(x, y)$ of each pixel during the classifier training. The rule aims to give less importance to pixels that may be not well binarized. After several erosions, if a pixel stays black or white then it is more likely to be well classified. At the opposite, if after one erosion, a black pixel becomes white then there are chances that it has not been well binarized. For an image I, lets consider the binarization of I, I_b. With I_b, we apply 3 successive erosions in order to get 3 new images. If a pixel remains black in the thrid image or if a pixel is white in all the 3 images then we strength of the pixel is full. It is considered as well labeled. The 3 erosions allow giving an estimation of the binarization model. For each pixel, we define a fixed strength for the different possible cases (1,0.8 and 0.6).

6 Experiments

We experimented the filtering method on ancient documents; they propose a challenge for all OCR. As opposed to modern-day documents, ancient documents can be divided in many categories, depending on their ages and their qualities. The documents set is composed in images extracted from a French journal of the XIXth century. In these documents, OCR may be used after some pre-processing. The processed images are compressed in JPEG. They are estimated to be 96dpi. Unfortunately, such bad quality may disable the recognition in some way. The actual challenge is to find a way to recover from this document issues in order to recognize the text. For the evaluation of the binarization and filtering methods on the recognition, we have used the commercial OCR Omnipage 14 and the raw FineReader 7 OCR engine [1, 2]. Commercial OCR have their own pre-processing algorithm. The comparison between OCR leads to give information about the pre-processing invariance against the inner OCR pre-processing.

For the OCR recognition, we did compare the DCS method with other classical binarization methods for documents: Bernsen, Niblack, Otsu and Sauvola. With the Niblack binarization, the OCR recognition is almost impossible as the method also allows us the text recognition of the back page. The Bernsen binarization lets many artifacts on the images; some characters are well segmented whereas many parts remain noisy. In the figure 2, the OCR results are presented for the best methods. Although commercial OCR possesses pre-processing modules, the binarized pages with the Sauvola method allows us a better recognition rate than the original image for both OCR. The table 2 presents an extract of a document page and the binarization and filtering effects. The image DCS (classifiers selection) illustrates the different pixel clusters, one for each gray level ($K = 100$). Compared to the classical binarization techniques in the table 1, the gain of the DCS method varies from 0,16% for Finereader7 and 1,06% for Omnipage14. Although, the difference on the images between the Sauvola and the DCS method is low, the impact on the OCR recognition rate shows the interest of the DCS method.

7 Conclusion

An efficient dynamic classifier selection method has been proposed. It is possible to outperform binarization techniques, thanks to a local filtering method with a dynamic classifier selection strategy. The method also provides better results than the considered ground truth for the training. Further works will include a feedback from the OCR results analysis as it could add information to where pre-processing methods must be applied.

References

[1] Finereader 7, abbyy, http://www.abbyy.com.

[2] Omnipage 14, scansoft, http://www.scansoft.com.

[3] C. H. Anderson and V. E. D.C. Shifter circuits: A computational strategy for directed visual attention and other dynamic neural processes. *Proc. Natl. Acad. of Sci. USA 84*, pages 6297–6301, 1987.

[4] G. Giacinto and F. Roli. Adaptive selection of image classifiers. *Proc. of the 9th ICIAP - LNCS 1310*, pages 38–45, 1997.

[5] M. Gupta, N. Jacobson, and E. Garcia. Ocr binarization and image pre-processing for searching historical documents. *Pattern Recognition, vol. 40, no. 2*, pages 389–397, 2007.

[6] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop, in neural networks: Tricks of the trade. *(G. Orr and Muller K., eds.)*, 1998.

[7] M. Levine and A. Nazif. Dynamic measurement of computer generated image segmentation. *IEEE PAMI, vol. 7, no. 2*, pages 155–164, 1985.

[8] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans, Syst Man Cybern, vol. 9, no. 1*, pages 62–66, 1979.

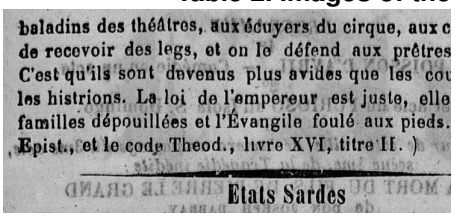
[9] G. Seni, V. Kripasundar, and R. Srihari. Generalizing edit distance for handwritten text recognition. *In Proceedings of SPIE/IS &T Conference on Document Recognition*, pages 54–65, 1995.

[10] S. Tabbone and L. Wendling. Multi-scale binarization of images. *Pattern Recognition Letters, vol. 24, no 1-3*, pages 403–411, 2003.

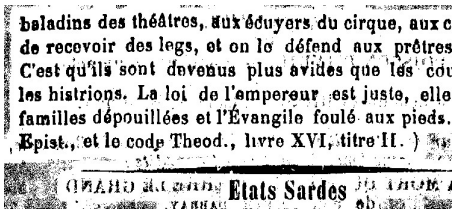
Table 1. OCR recognition.

	Recognition	Rejection	Confusion	Addition	Deletion	Cut	Fusion
Omnipage 14							
Original	88.04	0.00	2.14	5.78	3.43	0.38	0.21
Sauvola	93.40	0.00	3.21	1.37	1.21	0.56	0.26
Otsu	76.95	0.01	6.57	7.38	7.13	1.00	0.96
DCS + Kmeans	94.46	0.02	2.13	1.57	0.89	0.53	0.22
FineReader 7							
Original	20.44	0.50	8.54	13.75	54.61	1.07	1.08
Sauvola	82.16	0.55	7.73	3.45	3.60	1.66	0.85
Otsu	78.95	0.59	7.93	4.33	5.48	1.69	1.03
DCS + Kmeans	82.42	0.52	7.99	4.25	2.63	1.31	0.88

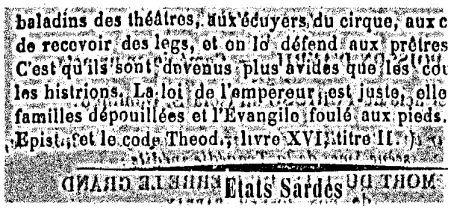
Table 2. Images of the journal "Le Petit Niois".



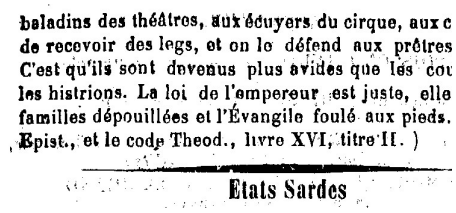
Original



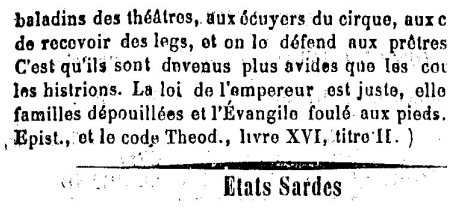
Bernsen



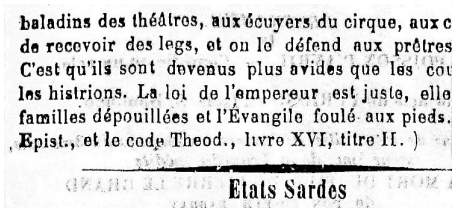
Niblack



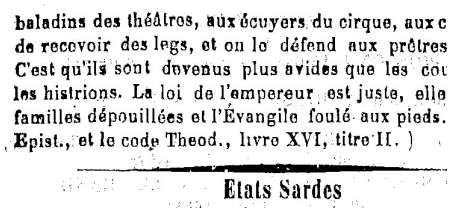
Otsu



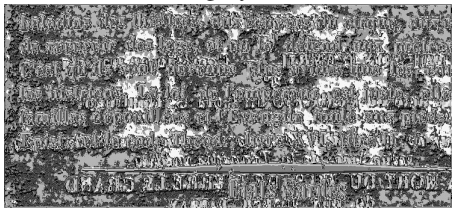
Sauvola



DCS (gray level)



DCS (binarization)



DCS (Classifiers selection)

[11] D. Trier and T. Taxt. Evaluation of binarization methods for document images. *IEEE PAMI, vol. 17, no. 3*, pages 312–315, 1995.